The current structural glycome landscape and emerging technologies

Liviu Copoiu[1] and Sony Malhotra[2]

[1]Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1GA, United Kingdom

[2]Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom

Corresponding author: Sony Malhotra

Email: s.malhotra@cryst.bbk.ac.uk

Keywords:

Protein-carbohydrate complexes, glycobiology, glycan-binding proteins, glycosylation, database, molecular dynamics.

Highlights

- Review the challenges to decipher the glycocode.
- Tools and repositories for investigating the protein-carbohydrate complexes.
- Experimental and computational methods for studying protein-carbohydrate interactions and to gain insights into dynamics and understand conformational changes.
- Recent molecular and cellular biology techniques (such as gene editing) and glycan engineering are very useful to understand the glycome.

Abstract

Carbohydrates represent one of the building blocks of life, along with nucleic acids, proteins and lipids. Although glycans are involved in a wide range of processes from embryogenesis to protein trafficking and pathogen infection, we are still a long way from deciphering the glycocode. In this review, we aim to present a few of the challenges that researchers working in the area of glycobiology can encounter and what strategies can be utilised to overcome them. Our goal is to paint a comprehensive picture of the current saccharide landscape available in the Protein Data Bank (PDB). We also review recently updated repositories relevant to the topic proposed, the impact of software development on strategies to structurally solve carbohydrate moieties, and state-of-the-art molecular and cellular biology methods that can shed some light on the function and structure of glycans.

Introduction

Compared to other essential biological building blocks such as nucleotides and proteins/peptides, the complexity of saccharides, as a class, is far greater. This is due to the variety of monosaccharide diversity, coupling, branching and anomeric states, as well as their ability to form free glycans or glycoconjugates (*e.g.* glycosylation, gangliosides). Furthermore, cell surface and secreted glycans present the lowest evolutionary conservation and exhibit the highest informational and structural diversity [1••]. Efforts to elucidate the combinatorial conundrum posed by the versatility of sugars have been purely theoretical, relying on approximations. The most recent number of putative combinations for hexa-saccharides is estimated to be 1.9×10^{11}, taking into account precise restrictions based on the chemical capacities of ten mammalian monosaccharides and anomeric expansion [2]. This number is several orders of magnitude larger than estimates for hexa-nucleic acids (4096 combinations) and hexa-peptides (6.4×10^{7} combinations).

One of the obstacles in deciphering the glycocode is the lack of a template from which the end-product glycans can be derived, unlike proteins that are coded by mRNA. Instead, glycan structures result from a complex interplay between partners that vary between cell types and cellular compartments. This complex, variable biological machinery gives rise to further complexity through "microheterogeneity", where a protein within the same cell/tissue can exhibit different glycoforms (glycosylation motifs), and "macroheterogeneity", where a protein can be glycosylated differently in different tissues [3].

Analysis of three-dimensional glycan-protein interactions, whether inter- or intra-molecular, could provide insights into the issues described above. Currently, there are ~5000 protein-carbohydrate (non-covalently bound) complexes (as reported in ProCarbDB [4•]) and ~7000 glycosylated structures (as reported in GlycanReader [5•]) in the Protein Data Bank [6]. However, analysing carbohydrate-containing structures is a non-trivial task, partly due to their complex and branched nature, as well as functional modifications of the monomeric units. In a recent review, Joosten and Lutteke [7•] provide a comprehensive picture of the challenges that structural biochemists must overcome in the field of glycoscience. The most important of these challenges are briefly listed below:

1) Not all saccharide moieties are correctly annotated in the PDB [8,9].
2) Significant errors in the structural rendering, partially due to incomplete electron density maps [7•,9], especially missing or incorrect glycosidic bonds.

3) Ligands might not be in the biologically relevant environment (*i.e.* missing aliphatic chains or acting as crystallographic additives bearing no biological role) and thus protein-carbohydrate interfaces cannot be correctly characterized [4•,10•].
4) Presence of non-host glycosylation motifs based on the expression system's carbohydrate repertoire [11•].
5) Scarcity of available glycan chains in the PDB, under 10% [5•], although more than 50% of proteins are estimated to present some level of glycosylation [12].

New repositories in glycobiology

Currently, there are issues regarding saccharide annotation (*e.g.* GlycosciencesDB.DE [13] identifies 776 monosaccharides while the PDB recognizes less than 600), as well as validation issues with carbohydrate moieties in the PDB. Thus, researchers have curated structural [4,5,14••–19••] and annotation [18–23] databases (Table 1) specifically for glycan binding proteins (GBPs) and saccharide containing structures. Unilectin3D (last updated in 2019) hosts experimentally solved structures for lectins, across all kingdoms (including viruses) [15]. Enzymes involved in carbohydrate synthesis or breakdown are extensively covered in CaZy (Carbohydrate-active enzyme) database [19]. The authors have mapped all possible structures from the PDB to their enzyme nomenclature. In doing this they have also, identifed over 100 types of carbohydrate-like molecules as biologically relevant ligands. Another useful online resource for glycan structures and motifs is GlyTouCan [14], which hosts over 100,000 carbohydrate structures and identifies 800 monosaccharides. The recently developed repository ProCarbDB identifies all protein-carbohydrate (non-covalently bound) complexes including those of enzymes, lectins, antibodies and transporter proteins [4]. Resources combining glycan structural information with data from nuclear magnetic resonance (NMR) experiments have also been developed in recent years: Carbohydrate Structure Database (CSDB) [17] and Glyco 3D [16]. Glycan Reader (last updated in 2018) is another useful resource which can automatically identify and annotate glycans within a PDB file, recognize glycosidic linkages and set up the glycan-protein complex for simulation using the CHARMM force field [5•]. Finally, DAGR (Database of Anti-Glycan Reagents) is a very useful repository for molecular biologists, because it offers a glycan IUPAC text search in order to identify the best antibody for a particular motif [23].

In order to provide a comprehensive picture of the available structural information for saccharide containing structures and apo-GBPs, we analysed the four largest and recently

updated datasets: Glycan Reader, CAZy, Unilectin3D and ProCarbDB. We have identified overlaps between these datasets as well as specific knowledge domains (Figure 1A). Unilectin3D and CAZy are hosting both apo and glycan bound forms of lectins and carbohydrate enzymes, respectively. Thus, it is not surprising that the majority (over 90%) of the unique entries (not overlapping with any other dataset) for both Unilectin3D and CAZy do not have a ligand. Glycan Reader identifies all carbohydrate moieties (including chemical modifications) present in the PDB, based on automated recognition of carbohydrates. In the subset of its unique entries, 68% are glycosylated non-GBPs. ProCarbDB hosts only protein-carbohydrate complexes, and thus in theory all its entries should be part of the Glycan Reader dataset. However, there are some ProCarbDB entries which are not present in the Glycan Reader dataset (such as 5M63 and 6B0K). This is due to the larger saccharide dictionary (merged set between PDB and pdb-care annotations) used in ProCarbDB. Another set of unique structures present only in Glycan Reader are those that have been crystallised using glycan containing surfactants compounds, such as β-octylglucoside (PDB Ligand ID: BOG), which is a crystallographic additive in the PDB structure 5ML5 and has no reported biological role. The divergence between datasets can be explained based on specificity, definition of carbohydrate moieties, biological function (or lack thereof) of the ligand and presence of obsolete structures.

Due to the presence of unique entries in each of the four repositories mentioned above, we merged everything into a unique set of 20,505 PDB structures and mapped each structure to its date of release (Figure 1B). The percentage per year of released structures either containing a saccharide moiety or describing a GBP is more or less constant (between 10-15%) for the last 10 years. Most of the structures, (19642 - 95.7%), present in the above-mentioned set were solved by X-ray crystallography, followed by electron microscopy, (323 - 1.6%), solution NMR, (241 - 1.2%) and other techniques, (69 - 0.3%). The remaining structures, (230 - 1.2%) are obsolete structures still present to some extent in all the datasets with the exception of ProCarbDB, which presently has no obsolete structures. Interestingly, the number of crystal structures solved by electron microscopy has steadily increased over the past 6 years: from only 4 structures containing saccharide moieties solved in 2014 to 122 in 2018 and already 83 by August 2019.

Structures containing glycans

The intrinsic flexibility of glycosidic linkages makes obtaining crystals for X-ray experiments a difficult and a time-consuming task. Other techniques such as NMR and computational techniques, including mathematical modelling and molecular dynamics (MD), might provide insights into different conformations of GBPs [24,25].

NMR approaches were used, for example, to define protein-glycan interfaces and the affinity for heparin of proMMP-7 (matrix metalloproteinase-7 zymogen) complex in solution [26]. Recently, several structural approaches (such as NMR, X-ray crystallography and cryo-EM) supported by MD simulations have been used to shed light on dynamic processes occurring in bacterial membranes such as: conformational preference and dynamics of O-antigen in *E. coli* [27], processing and elongation of peptidoglycans [28], biological and biophysical behaviour of bacterial membranes [29], protein-outer membrane interactions [30–32], and sugar transport [33•].

Other fields benefiting from the power of MD simulations are pathology and immunology, especially in areas concerning: antigenicity and immunogenicity [34–37], and identification of novel structural motifs [38,39]. The interactions between interleukin-10 (IL-10) and its ligand glycosaminoglycans (GAG) have been defined using X-ray crystallography (apo IL-10), NMR (IL-10-GAG complex), molecular docking and MD simulations [40]. In an another example, small angle X-ray scattering and mathematical modelling were used characterise CC chemokine ligand (CCL) 5 oligomerisation process and unveil the interactions of CCL5 an CCL3 with GAG [41]. Recently, a 3.6Å cryo-EM structure of the icosahedral capsid of the porcine circovirus 2 complex with heparin was solved and identified the non-symmetrical distribution of heparin to the icosahedral virus [42]. Cryo-EM has also been used to solve the structures of membrane bound proteins such as ATP binding cassette transporters (comprising of lipopolysaccharides transport protein subunits) in bacteria [43••].

The N-glycosylation landscape was analysed by Lee *et al*. and statistically no significant global conformational change was recorded between glycosylated and de-glycosylated forms of the same protein [44•]. However, this result contradicts a previous study by Xin and Radivojac [45]. Lee *et al.,* performed 200-ns MD simulations that showed decreased dynamics upon glycosylation, not at the glycosylation site itself, but allosterically in other regions. Suga *et al*. performed a different holistic analysis on the N-glycosylation available in the PDB focusing on the processing and maturation of glycosylation [11•]. They found that the accessible surface

area (ASA) of immaturely glycosylated (only the early steps in glycosylation) asparagine is lower than that of mature glycans. The authors conducted a protein surface analysis around the N-glycosylation sites and suggested that there is an overall bias for γ-branched amino acids (Asn, Asp and Leu) distribution around the immature N-glycans. However, these analyses use the currently available data, which are limited. Furthermore, we can only inspect the glycan patterns of the fully folded proteins.

Although three-dimensional glycosylation data are limited and biased, a number of bioinformatic tools were developed recently for predicting [46], validating [47••], identifying [5,9], modelling [48–53] and simulating [49••,54] structures that contain glycan moieties (Table 2). The choice of modelling software depends on the experimental details and MD simulation program used. However, all bioinformatics tools presented in Table 2 are reported to generate three-dimensional glycan structures with good confidence (*i.e*. low root mean square deviation (RMSD) between the modelled and the experimentally generated glycan three-dimensional structures).

There are limitations of current approaches and usually each software has a different nomenclature scheme for encoding carbohydrates, which makes it difficult for the users to utilise multiple programs with one protocol. Furthermore, only CHARMM-GUI set of tools can be used with all major MD simulation programs, which further limits experimental design.

Protein-carbohydrate interactions

Protein-carbohydrate interaction partners can be identified by using high-throughput assays such as glycan arrays [55] or lectin arrays [56]. Manual interpretation of the results is a very laborious and time-consuming task. However, a couple of recently developed algorithms (*e.g.* GlycoSearch [57] and MACAW (Multiple Carbohydrate Alignment with Weights) [58]) can automatically analyse glycan array data.

As explained in the introduction of this review, comprehending the diversity of glycans has been achieved largely using theoretical approaches and the structural information available is limited and prone to mistakes. Thus, any currently viable algorithm developed either for statistical analysis or prediction of protein-carbohydrate interactions is highly specific for one type of GBP. Samsonov *et al*. have analysed protein-glycosaminoglycan (GAG) interactions using MD simulations and uncovered interactions between GAG processing enzymes and their ligands that are less electrostatics driven than interactions between GAG binding proteins

(lacking enzymatic activity) and their ligands. Furthermore, they proved that Poisson–Boltzmann surface area (PBSA)-based electrostatic potential calculations are robust enough to predict putative GAG-binding regions [59].

With the revolution of machine learning at the beginning of this decade, a couple of algorithms have also been developed in the field of glycobiology. Nassif *et al*. tried to predict protein-glucose binding sites using support vector machines (SVM) [60]. They used random forest (RF) as a feature selection tool and identified critical variables for accurate predictions including charge, hydrogen bonding ability, hydrophobicity, type of residues as well as the presence of ordered water molecules and ions present in the X-ray crystal structure. The surprising result is that amongst all chemical properties described above both hydrophobicity and charge are more important than hydrogen bonding ability. Their algorithm achieved 8.11% error, 89.66% sensitivity and 93.33% specificity over their dataset.

A more recent article by Pai *et al*. aimed to predict protein-mannose interacting residues using both SVM and RF [61]. They identified position-specific scoring matrix (PSSM) as being a better feature set than local amino acid composition for predicting mannose-interacting residues. They report a significant improvement in specificity, sensitivity and precision when compared to other state-of-art predictors.

Although these results show good results for very specific tasks it is important to point out one caveat: any machine learning program requires accurate and large training sets, devoid of false positive entries and duplicates, in order to generate accurate predictions. Nevertheless, by carefully analysing each individual dataset, we were able to identify false positive entries. The degree to which those entries skew the results is outside the scope of this article.

Recent molecular and cellular biology tools to investigate the glycome

The current structural information on carbohydrate moieties and motifs as a whole is sparse, not aggregated, non-uniform and sometimes even conflicting. At the same time our knowledge about the importance of glycans in protein uptake, trafficking and localization is opening new avenues for the generation of highly specific therapeutic agents using carbohydrates, and their analogues [1]. Since there is little structural information available, generating controlled glycan-patterns can be used as an alternative to probe binding pockets for GPBs [62••,63].

With the engineering of CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) associated protein 9 (CAS9) and zinc finger nucleases (ZFN), targeted and refined cell-based

assays are now available in glycobiology [64]. Basic research has harvested the power of these molecular tools to:

1) Generate a guide RNA (gRNA) library targeting all known glycosyltransferases in HEK297 cells (Human Embryonic Kidney cells) [65•].
2) Control-synthesize a large part of the human glycome [66].
3) Understand the redundancy and specificity of O-glycosylation pathways [67].
4) Generate specific monoclonal antibodies against mucin-type O-glycosylation [68•].

Translational research has also started to focus on glycan engineering by:

1) Generating for the first time homogenous glycosylation patterns, thus alleviating "microheterogeneity" [69].
2) Improving both circulation time and tissue specificity of proteins based on different glycosylation patterns [70••].
3) Identifying carbohydrate analogues with higher affinity against putative targets using novel cell based glycan arrays, that better simulate the *in vivo* environment [62••].

We would like to point the readers to the book: *"Essentials of Glycobiology"*, chapters 56-60 [71], which very well explains the history of methods, techniques and therapeutics developed using glycosylation engineering. We would like to emphasize the rapid rhythm at which new, state-of-the-art biological tools in the realm of glycobiology are being generated. We know that different glycosylation patterns will affect: the molecular weight, charge and solubility of a protein, however, how to control these properties was until recently a conundrum. It has been shown [66,69,70••] that precise glycoengineering is a very complex process requiring both removing certain genes (knock-down) and introducing new ones (knock-in). In order to know which proteins to combine to achieve the desired motif, a complete dictionary of glycosyltransferases [19••], their functional dependencies [67] and how to target them [65•] is required. This type of deconstruction of the glycosylation atlas allows researchers to inspect biological functions of specific glycan motifs such as: circulation time and organ specificity for enzyme replacement therapy [70••], efficiency and safety of erythropoietin [72] and osteoclast formation [62••].

Even though early glycoengineering efforts used a structural based approach for developing zanamivir [73] and oseltamivir [74] against influenza virus, much of our current knowledge about the glycocalyx and host-pathogen interactions is disjunct from structural biochemistry.

The only way to characterize protein-glycan interfaces at the atomic level, is to have accurate structural information about glycans.

Conclusions

We now know that almost every human disease involves glycans at a step in its evolution [75]. Our knowledge of glycan-binding proteins is increasing at a very fast pace, however, carbohydrates still remain "the dark matter of biology". This review focuses on the current structural glycome landscape and its shortcomings. One caveat of the carbohydrate-related structural information is the fact that it is highly prone to human mistakes. The main emphasis of this article is that we need to rethink and redesign the way we store the structural data of glycans. Looking forward, we will need to generate exhaustive and curated datasets of structures containing carbohydrates to harvest the full potential of emerging research techniques, such as deep learning. Furthermore, combining computational biology, structural biochemistry, molecular biology and genetics will be crucial in order to achieve a better understanding of glycobiology and provide a uniform stream of knowledge.

Acknowledgements

We would like to thank Prof. Sir Tom L. Blundell for reviewing the text and providing his suggestions. We would also like to thank Dr. Pedro Henrique Monteiro Torres for valuable inputs. Furthermore, we thank Dr Wonpil Im for generating an up-to-date dataset of Glycan Reader, Dr. Bernard Herissat for providing the Cazy DB dataset, and the authors of Unilectin 3D for making their dataset available. We thank Dr Elena Fonfira and Dr Sai Man Liu from Ipsen Bioinnovation Ltd. for their expertise, suggestions and useful discussions.

Figure Legends:

Figure 1: Structural statistics of carbohydrate moieties and glycan binding proteins. A) Venn diagram of four largest and their respective last updated datasets. B) Percentage per year of released structures in the protein databank in each of the five classes marked in the legend.

References

Articles of interested have been highlighted as:

• of special interest

•• of outstanding interest

1••. Varki A: **Biological roles of glycans**. *Glycobiology* 2017, **27**:3–49.

A comprehensive review on past and current knowledge of glycans, taking into account structural and functional roles of carbohydrates.

2. Werz DB, Ranzinger R, Herget S, Adibekian A, von der Lieth C-W, Seeberger PH: **Exploring the structural diversity of mammalian carbohydrates ("glycospace") by statistical databank analysis.** *ACS Chem Biol* 2007, **2**:685–91.

3. Broussard AC, Boyce M: **Life is sweet: The cell biology of glycoconjugates**. *Mol Biol Cell* 2019, **30**:525–529.

4•. Copoiu L, Torres PHM, Ascher DB, Blundell TL, Malhotra S: **ProCarbDB: a database of carbohydrate-binding proteins.** *Nucleic Acids Res* 2019.

Database of protein-carbohydrate complexes hosting a wide variety of glycan binding proteins: from lectins to antibodies and transporter proteins.

5•. Park S, Lee J, Patel DS, Ma H, Lee HS, Jo S, Im W: **Structural bioinformatics Glycan Reader is improved to recognize most sugar types and chemical modifications in the Protein Data Bank**. 2017, **33**:3051–3057.

Large repository identifing glycan structures in PDB structures. Useful statistics about glycans present in PDB.

6. Burley SK, Berman HM, Bhikadiya C, Bi C, Chen L, Di Costanzo L, Christie C, Dalenberg K, Duarte JM, Dutta S, et al.: **RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy**. *Nucleic Acids Res* 2019, **47**:D464–D474.

7•. Joosten RP, Lütteke T: **Carbohydrate 3D structure validation**. *Curr Opin Struct Biol* 2017, **44**:9–17.

Comprehensive review on current issues with structurally solving carbohydrate moieties as well as possible ways to rectify erroneous glycan structures.

8. Lütteke T, Frank M, von der Lieth C-W: **Data mining the protein data bank: automatic detection and assignment of carbohydrate structures.** *Carbohydr Res* 2004, **339**:1015–20.

9. Lütteke T, von der Lieth C-W: **pdb-care (PDB carbohydrate residue check): a program to support annotation of complex carbohydrate structures in PDB files.** *BMC Bioinformatics* 2004, **5**:69.

10•. Hamark C, Berntsson RP-A, Masuyer G, Henriksson LM, Gustafsson R, Stenmark P, Widmalm G: **Glycans Confer Specificity to the Recognition of Ganglioside Receptors by Botulinum Neurotoxin A.** *J Am Chem Soc* 2017, **139**:218–230.

Study combining structural approaches with biophysical assays in order to quantitatively characterize botulinum toxin's affinity towards gangliosides

11•. Suga A, Nagae M, Yamaguchi Y: **Analysis of protein landscapes around N -glycosylation sites from the PDB repository for understanding the structural basis of N -glycoprotein processing and maturation**. 2018, **28**:774–785.

Large scale analysis of N-glycosylation chains present in the PDB provinding evidance that the amino acid composition around glycosylation sites is biased for immature glycoproteins. Furthermore, they found out that solvent accessibility area around immature glycans is significantly lower than around mature glycans.

12. Apweiler R: **On the frequency of protein glycosylation , as deduced from analysis of the SWISS-PROT database**. 1999, **1473**:4–8.

13••. Böhm M, Bohne-Lang A, Frank M, Loss A, Rojas-Macias MA, Lütteke T: **Glycosciences.DB: an annotated data collection linking glycomics and proteomics data (2018 update)**. *Nucleic Acids Res* 2019, **47**.

Large repository for glycan structures containing the highest monosaccharide dictionary.

14•. Tiemeyer M, Aoki K, Paulson J, Cummings RD, York WS, Karlsson NG, Lisacek F, Packer NH, Campbell MP, Aoki NP, et al.: **GlyTouCan: An accessible glycan structure repository**. *Glycobiology* 2017, **27**:915–919.

Periodically updated structural repository identifing lectins in PDB.

15. Bonnardel F, Mariethoz J, Salentin S, Robin X, Schroeder M, Perez S, Lisacek FDS, Imberty A: **Unilectin3d, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands**. *Nucleic Acids Res* 2019, **47**:D1236–D1244.

16. Pérez S, Sarkar A, Rivet A, Breton C, Imberty A: **Glyco3D: a portal for structural glycosciences.** *Methods Mol Biol* 2015, **1273**:241–58.

17. Toukach P V., Egorova KS: **Carbohydrate structure database merged from bacterial, archaeal, plant and fungal parts**. *Nucleic Acids Res* 2016, **44**:D1229–D1236.

18. Choudhary P, Nagar R, Singh V, Bhat AH, Sharma Y, Rao A: **ProGlycProt V2.0, a repository of experimentally validated glycoproteins and protein glycosyltransferases of prokaryotes**. *Glycobiology* 2019, **29**:461–468.

19••. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B: **The carbohydrate-active enzymes database (CAZy) in 2013.** *Nucleic Acids Res* 2014, **42**:D490-5.

Novel nomenclature derived specifically for carbohydrate processing enzymes. It identifies over 100 distinct carbohydrate-like structures in PDB structures.

20. Terrapon N, Lombard V, Drula É, Lapébie P, Al-Masaudi S, Gilbert HJ, Henrissat B: **PULDB: the expanded database of Polysaccharide Utilization Loci.** *Nucleic Acids Res* 2018, **46**:D677–D683.

21. Birch J, Van Calsteren M-R, Pérez S, Svensson B: **The exopolysaccharide properties and structures database: EPS-DB. Application to bacterial exopolysaccharides.** *Carbohydr Polym* 2019, **205**:565–570.

22. Clerc O, Deniaud M, Vallet SD, Naba A, Rivet A, Perez S, Thierry-Mieg N, Ricard-Blum S: **MatrixDB: integration of new data with a focus on glycosaminoglycan interactions.** *Nucleic Acids Res* 2019, **47**:D376–D381.

23. Sterner E, Flanagan N, Gildersleeve JC: **Perspectives on Anti-Glycan Antibodies Gleaned from Development of a Community Resource Database.** *ACS Chem Biol* 2016, **11**:1773–83.

24. Wormald MR, Petrescu AJ, Pao Y-L, Glithero A, Elliott T, Dwek RA: **Conformational studies of oligosaccharides and glycopeptides: complementarity of NMR, X-ray crystallography, and molecular modelling.** *Chem Rev* 2002, **102**:371–86.

25. Fadda E, Woods RJ: **Molecular simulations of carbohydrates and protein-carbohydrate interactions: motivation, issues and prospects.** *Drug Discov Today* 2010, **15**:596–609.

26. Fulcher YG, Prior SH, Masuko S, Li L, Pu D, Zhang F, Linhardt RJ, Van Doren SR: **Glycan Activation of a Sheddase: Electrostatic Recognition between Heparin and proMMP-7.** *Structure* 2017, **25**:1100-1110.e5.

27. Blasco P, Patel DS, Im W: **Conformational Dynamics of the Lipopolysaccharide from Escherichia coli O91 Revealed by Nuclear Magnetic Resonance Spectroscopy and Molecular Simulations**. 2017, doi:10.1021/acs.biochem.7b00106.

28. Kim S, Pires MM, Im W: **Insight into Elongation Stages of Peptidoglycan Processing in Bacterial Cytoplasmic Membranes**. *Sci Rep* 2018, doi:10.1038/s41598-018-36075-y.

29. Hughes A V., Patel DS, Widmalm G, Klauda JB, Clifton LA, Im W: **Physical Properties of Bacterial Outer Membrane Models: Neutron Reflectometry & Molecular Simulation**. *Biophys J* 2019, **116**:1095–1104.

30. Lee J, Patel DS, Kucharska I, Tamm LK, Im W: **Refinement of OprH-LPS Interactions by Molecular Simulations**. *Biophys J* 2017, **112**:346–355.

31. Lee J, Pothula KR, Im W: **Simulation Study of Occk5 Functional Properties in Pseudomonas aeruginosa Outer Membranes**. 2018, doi:10.1021/acs.jpcb.8b07109.

32. Mobarak E, Håversen L, Manna M, Rutberg M, Levin M, Perkins R, Rog T, Vattulainen I, Borén J: **Glucosylceramide modifies the LPS-induced inflammatory response in macrophages and the orientation of the LPS/TLR4 complex in silico.** *Sci Rep* 2018, **8**:13600.

33•. Ren Z, Lee J, Moosa MM, Nian Y, Hu L, Xu Z, McCoy JG, Ferreon ACM, Im W, Zhou M: **Structure of an EIIC sugar transporter trapped in an inward-facing conformation**. *Proc Natl Acad Sci* 2018, **115**:5962–5967.

A study combining MD simulation and X-ray crystallography providing evidence that sugar translocation can be achieved by movement of a carbohydrate-binding motif.

34. Kuttel MM, Ravenscroft N: **Conformation and Cross-Protection in Group B Streptococcus Serotype III and Streptococcus pneumoniae Serotype 14: A Molecular Modeling Study.** *Pharmaceuticals (Basel)* 2019, **12**.

35. Urbanowicz RA, Wang R, Schiel JE, Keck Z-Y, Kerzic MC, Lau P, Rangarajan S, Garagusi KJ, Tan L, Guest JD, et al.: **Antigenicity and Immunogenicity of Differentially Glycosylated Hepatitis C Virus E2 Envelope Proteins Expressed in Mammalian and Insect Cells.** *J Virol* 2019, **93**.

36. Havenar-Daughton C, Sarkar A, Kulp DW, Toy L, Hu X, Deresa I, Kalyuzhniy O, Kaushik K, Upadhyay AA, Menis S, et al.: **The human naive B cell repertoire contains distinct subclasses for a germline-targeting HIV-1 vaccine immunogen.** *Sci Transl Med* 2018, **10**.

37. Amon R, Grant OC, Leviatan Ben-Arye S, Makeneni S, Nivedha AK, Marshanski T, Norn C, Yu H, Glushka JN, Fleishman SJ, et al.: **A combined computational-experimental approach to define the structural origin of antibody recognition of sialyl-Tn, a tumor-associated carbohydrate antigen.** *Sci Rep* 2018, **8**:10786.

38. Kuttel MM, Cescutti P, Distefano M, Rizzo R: **Fluorescence and NMR spectroscopy together with molecular simulations reveal amphiphilic characteristics of a Burkholderia biofilm exopolysaccharide.** *J Biol Chem* 2017, **292**:11034–11042.

39. Azurmendi HF, Battistel MD, Zarb J, Lichaa F, Negrete Virgen A, Shiloach J, Freedberg DI: **The β-reducing end in α(2-8)-polysialic acid constitutes a unique structural motif.** *Glycobiology* 2017, **27**:900–911.

40. Künze G, Köhling S, Vogel A, Rademann J, Huster D: **Identification of the Glycosaminoglycan Binding Site of Interleukin-10 by NMR Spectroscopy.** *J Biol Chem* 2016, **291**:3100–13.

41. Liang WG, Triandafillou CG, Huang T-Y, Zulueta MML, Banerjee S, Dinner AR, Hung S-C, Tang W-J: **Structural basis for oligomerization and glycosaminoglycan binding of CCL5 and CCL3.** *Proc Natl Acad Sci U S A* 2016, **113**:5000–5.

42. Dhindwal S, Avila B, Feng S, Khayat R: **Porcine Circovirus 2 Uses a Multitude of**

Weak Binding Sites To Interact with Heparan Sulfate, and the Interactions Do Not Follow the Symmetry of the Capsid. *J Virol* 2019, **93**.

43••. Li Y, Orlando BJ, Liao M: **Structural basis of lipopolysaccharide extraction by the LptB2FGC complex.** *Nature* 2019, **567**:486–490.

Single-particle cryo-electron microscopy study providing evidence for a role of LptC in lipopolysaccharide transport, by interacting with $LptB_2FG$.

44. Lee HS, Qi Y, Im W: **Effects of N -glycosylation on protein conformation and dynamics : Protein Data Bank analysis and molecular dynamics simulation study**. 2015, doi:10.1038/srep08926.

45. Xin F, Radivojac P: **Post-translational modifications induce significant yet not extreme changes to protein structure.** *Bioinformatics* 2012, **28**:2905–13.

46. Taherzadeh G, Dehzangi A, Golchin M, Zhou Y, Campbell MP: **SPRINT-Gly: predicting N- and O-linked glycosylation sites of human and mouse proteins by using sequence and predicted structural properties**. *Bioinformatics* 2019, doi:10.1093/bioinformatics/btz215.

47••. Agirre J, Iglesias-Fernández J, Rovira C, Davies GJ, Wilson KS, Cowtan KD: **Privateer: software for the conformational validation of carbohydrate structures**. *Nat Struct Mol Biol* 2015, **22**:833–834.

The article presents a software for automated conformational validation of carbohydrates, that is part of CCP4.

48••. Labonte JW, Adolf-Bryfogle J, Schief WR, Gray JJ: **Residue-centric modeling and design of saccharide and glycoconjugate structures.** *J Comput Chem* 2017, **38**:276–287.

This article introduces RossettaCarbohydrate framework which is a tool for modelling both glycans and glycoconjucates using a residue-centric approach.

49••. Park SJ, Lee J, Qi Y, Kern NR, Lee HS, Jo S, Joung I, Joo K, Lee J, Im W: **CHARMM-GUI Glycan Modeler for modeling and simulation of carbohydrates and glycoconjugates**. *Glycobiology* 2019, **29**:320–331.

This article introduces Glycan Modeler, which is built on top of Glycan Reader and enables

the generation of glycan and glycoconjugate structures providing valid input setup for all major MD simulation programs.

50. Kuttel MM, Ståhle J, Widmalm G: **CarbBuilder: Software for building molecular models of complex oligo- and polysaccharide structures.** *J Comput Chem* 2016, **37**:2098–105.

51. Baltoumas FA, Hamodrakas SJ, Iconomidou VA: **The gram-negative outer membrane modeler: Automated building of lipopolysaccharide-rich bacterial outer membranes in four force fields.** *J Comput Chem* 2019, **40**:1727–1734.

52•. Lee J, Patel DS, Ståhle J, Park S, Kern NR, Kim S, Lee J, Cheng X, Valvano MA, Holst O, et al.: **CHARMM-GUI Membrane Builder for Complex Biological Membrane Simulations with Glycolipids and Lipoglycans**. 2019, doi:10.1021/acs.jctc.8b01066.

The authors have developed a web-based platform that enables users to build complex biological structures such as lipopolysaccharides and lipooligosaccharides that can be embedded in biological membranes for simulation studies.

53. Singh A, Montgomery D, Xue X, Foley BL, Woods RJ: **GAG Builder: a web-tool for modeling 3D structures of glycosaminoglycans**. *Glycobiology* 2019, **29**:515–518.

54. Danne R, Poojari C, Martinez-seara H, Rissanen S, Lolicato F, Vattulainen I: **doGlycans − Tools for Preparing Carbohydrate Structures for Atomistic Simulations of Glycoproteins, Glycolipids, and Carbohydrate Polymers for GROMACS**. 2017, doi:10.1021/acs.jcim.7b00237.

55. Feizi T, Fazio F, Chai W, Wong CH: **Carbohydrate microarrays - a new set of technologies at the frontiers of glycomics.** *Curr Opin Struct Biol* 2003, **13**:637–45.

56. Pilobello KT, Krishnamoorthy L, Slawek D, Mahal LK: **Development of a lectin microarray for the rapid analysis of protein glycopatterns.** *Chembiochem* 2005, **6**:985–9.

57. Kletter D, Cao Z, Bern M, Haab B: **Determining lectin specificity from glycan array data using motif segregation and GlycoSearch software.** *Curr Protoc Chem Biol* 2013, **5**:157–69.

58. Hosoda M, Akune Y, Aoki-Kinoshita KF: **Development and application of an**

algorithm to compute weighted multiple glycan alignments. *Bioinformatics* 2017, **33**:1317–1323.

59. Samsonov SA, Pisabarro MT: **Computational analysis of interactions in structurally available protein-glycosaminoglycan complexes.** *Glycobiology* 2016, **26**:850–861.

60. Nassif H, Al-Ali H, Khuri S, Keirouz W: **Prediction of protein-glucose binding sites using support vector machines.** *Proteins* 2009, **77**:121–32.

61. Pai PP, Mondal S: **MOWGLI: prediction of protein-MannOse interacting residues With ensemble classifiers usinG evoLutionary Information.** *J Biomol Struct Dyn* 2016, **34**:2069–83.

62••. Briard JG, Jiang H, Moremen KW, MacAuley MS, Wu P: **Cell-based glycan arrays for probing glycan-glycan binding protein interactions**. *Nat Commun* 2018, **9**:1–11.

The authors have developed a fast method to develop a cell-based glycan assay that better simulates the in vivo environment than classical glycan arrays. Using this platoform they engineer high-affinity glycan ligands for Siglec-15.

63. Huang ML, Godula K: **Nanoscale materials for probing the biological functions of the glycocalyx.** *Glycobiology* 2016, **26**:797–803.

64. Steentoft C, Bennett EP, Schjoldager KTBG, Vakhrushev SY, Wandall HH, Clausen H: **Precision genome editing: A small revolution for glycobiology**. *Glycobiology* 2014, **24**:663–680.

65•. Narimatsu Y, Joshi HJ, Yang Z, Gomes C, Chen YH, Lorenzetti FC, Furukawa S, Schjoldager KT, Hansen L, Clausen H, et al.: **A validated gRNA library for CRISPR/Cas9 targeting of the human glycosyltransferase genome**. *Glycobiology* 2018, **28**:295–305.

Comprehensive study generating a comprehensive library of gRNA sequences that can be used to target human glycosyltransferase in order to better understand their impact in glycosylation.

66. Narimatsu Y, Joshi HJ, Nason R, Van Coillie J, Karlsson R, Sun L, Ye Z, Chen Y-H, Schjoldager KT, Steentoft C, et al.: **An Atlas of Human Glycosylation Pathways**

Enables Display of the Human Glycome by Gene Engineered Cells. *Mol Cell* 2019, **75**:394-407.e5.

67. Narimatsu Y, Joshi HJ, Schjoldager KT, Hintze J, Halim A, Steentoft C, Nason R, Mandel U, Bennett EP, Clausen H, et al.: **Exploring Regulation of Protein O-Glycosylation in Isogenic Human HEK293 Cells by Differential O-Glycoproteomics.** *Mol Cell Proteomics* 2019, **18**:1396–1409.

68•. Steentoft C, Yang Z, Wang S, Ju T, Vester-Christensen MB, Festari MF, King SL, Moremen K, Larsen ISB, Goth CK, et al.: **A validated collection of mouse monoclonal antibodies to human glycosyltransferases functioning in mucin-type O-glycosylation.** *Glycobiology* 2019, **29**:645–656.

The authors provide rigorous specificity validation for their monoclonal antibodies using human cell lines genetically engineered to express or not relevant glycosyltransferase.

69. Yang Z, Wang S, Halim A, Schulz MA, Frodin M, Rahman SH, Vester-Christensen MB, Behrens C, Kristensen C, Vakhrushev SY, et al.: **Engineered CHO cells for production of diverse, homogeneous glycoproteins**. *Nat Biotechnol* 2015, **33**:842–844.

70••. Tian W, Ye Z, Wang S, Schulz MA, Van Coillie J, Sun L, Chen YH, Narimatsu Y, Hansen L, Kristensen C, et al.: **The glycosylation design space for recombinant lysosomal replacement enzymes produced in CHO cells**. *Nat Commun* 2019, **10**:1–13.

A translational approach using gene engineering tools to increase the circulation time and organ specificity of α-galactosidase A.

71. Editors: Ajit Varki, Executive Editor, Richard D Cummings, Jeffrey D Esko, Pamela Stanley, Gerald W Hart, Markus Aebi, Alan G Darvill, Taroh Kinoshita, Nicolle H Packer, James H Prestegard, Ronald L Schnaar and PHS: *Essentials of Glycobiology*. Cold Spring Harbor Laboratory Press; 2017.

72. Čaval T, Tian W, Yang Z, Clausen H, Heck AJR: **Direct quality control of glycoengineered erythropoietin variants.** *Nat Commun* 2018, **9**:3342.

73. von Itzstein M, Wu WY, Kok GB, Pegg MS, Dyason JC, Jin B, Van Phan T, Smythe ML, White HF, Oliver SW: **Rational design of potent sialidase-based inhibitors of**

influenza virus replication. *Nature* 1993, **363**:418–23.

74. Kim CU, Lew W, Williams MA, Liu H, Zhang L, Swaminathan S, Bischofberger N, Chen MS, Mendel DB, Tai CY, et al.: **Influenza neuraminidase inhibitors possessing a novel hydrophobic interaction in the enzyme active site: design, synthesis, and structural analysis of carbocyclic sialic acid analogues with potent anti-influenza activity.** *J Am Chem Soc* 1997, **119**:681–90.

75. National Research Council (US) Committee on Assessing the Importance and Impact of Glycomics and Glycosciences.: *Transforming Glycoscience*. National Academies Press; 2012.

Table 1: Glycan repositories.

Name	Description	URL	Citation
GlyTouCan	Largest glycan structure repository	https://glytoucan.org/	Tiemeyer et al. 2017 [14•]
Unilectin 3D	Curated structures of carbohydrate binding proteins(lectins)	https://www.unilectin.eu/unilectin3D/	Bonnardel et al. 2019 [15]
Glyco 3D	Structural Database of lectins, mAb, and GAG-binding proteins	http://glyco3d.cermav.cnrs.fr/home.php	Pérez et al. 2015 [16]
CaZy	The carbohydrate-active enzymes database	http://www.cazy.org/	Lombard et al. 2014 [19••]
ProCarbDB	Database hosting protein-carbohydrate complexes	http://www.procarbdb.science/procarb/	Liviu el al. 2019 [4•]
CSDB	structural, taxonomic, NMR datafor bacterial, achaeal, plant and fungal	http://csdb.glycoscience.ru/gt.html	Toukach et al. 2016 [17]
ProGlycProt V2.0	Glycoproteins and glycosyltransferases of prokaryotes	www.proglycprot.org	Choudhary el al. 2019 [18]
PULDB	Bacteroidetes gram-negative bacteria Polysaccharide Utilization Loci annotation	http://www.cazy.org/PULDB_new/	Terrapon et al. 2018 [20]
EPS DB	Bacterial Exopolysaccharide Properties and Structures Database	http://www.epsdatabase.com/database.html	Birch et al. 2019 [21]
Matrix DB	Database with a focus on glycosaminoglycan interactions	http://matrixdb.univ-lyon1.fr/	Clerc et al. 2019 [22]
DAGR	Database of Anti-Glycan Reagents and Antibodies	https://ccr2.cancer.gov/resources/Cbl/Tools/Antibody/	Sterner et al. 2016 [23]

Table 2 Software for glycome prediction, modelling and validation.

Name	Description	URL	Citation
Glycan Reader	PDB Sugar identification and simulation preparation	http://glycanstructure.org/glycanreader/	Park et al. 2017 [5•]
RossetaCarbohydrate	Modeling and Design of Saccharide and Glycoconjugate Structures	https://www.rosettacommons.org	Labonte et al. 2017 [48••]
Glycan Modeller	Modelling N-/O-glycosylation on the target protein	http://glycanstructure.org/glycanmodeler	Park et al. 2019 [49••]
CarbBuilder	Models of carbohydrates from both mammalian and bacterial origin	https://people.cs.uct.ac.za/~mkuttel/Downloads.html	Kuttel et al.2016 [50]
doGlycans	Glycoprotein, glycolipids, carbohydrate polymers preparation for atomistic MD	https://pubs.acs.org/doi/10.1021/acs.jcim.7b00237	Danne et al. 2017 [54]
GNOMM	Automated building of lipopolysaccharide-rich bacterial outer membranes	http://bioinformatics.biol.uoa.gr/GNOMM	Baltoumas et al. 2019 [51]
LPS Modeller	Modelling lipopolysaccharides	http://charmm-gui.org/?doc=input/lps	Lee et al. 2019 [52•]
Glycolipid Modeller	Modelling glycolipids	http://charmm-gui.org/?doc=input/glycolipid	Lee et al. 2019 [52•]
GAG Builder	Modeling 3D structures of glycosaminoglycans	www.glycam.org/gag	Singh et al. 2019 [53]
Privateer	Conformational validation of carbohydrate structures	http://www.ccp4.ac.uk/html/privateer.htm	Agirre et al. 2015 [47••]
SPRINT-Gly	Seq prediction for N-/O-linked glycosylation sites of human and mouse	http://sparks-lab.org/server/SPRINT-Gly/	Taherzadeh et al. 2019 [46]